Report on Data Management and Data-Management Plans
for
The History of Science Society Committee on Research and the Profession

(September 6, 2013)

Task Force Members*

Daniel Goldstein (chair and principal author), *University of California, Davis*
Rafael Burgos-Mirabal, *University of Massachusetts, Amherst*
Cathryn Carson, *University of California, Berkeley*
David Caruso, *Chemical Heritage Society*
Matthew K. Chew, *Arizona State University*
Hamilton Cravens, *Iowa State University*
Julia Damerow, *Arizona State University*
Stephen DeRose, *Westminster Seminary*
David Grier, *George Washington University*
Manfred Laubichler, *Arizona State University*
Sarah Lowengard, *The Cooper Union*
Erika Milam, *Princeton University*
Erick Peirson, *Arizona State University*
James Skee, *University of California, Berkeley*
Virginia Trimble, *University of California, Irvine*
Brant Vogel, *Independent Scholar*

*Affiliation provided for identification purposes only.

CHARGE

"The Committee on Research and the Profession (CoRP) is forming an ad hoc task force to address the U.S. federal government's recent requirements concerning data management and data-management plans. Because federal agencies, particularly the National Science Foundation, define data by the standards of a given discipline, it is important that the History of Science Society have a sustained discussion both about what constitutes data for historians, and how best to share federally funded data with the broader public.

This short-term committee will draft a report for the profession's use that summarizes the problem, provides some examples of what might constitute data in various fields (including our own), and discusses advantages and disadvantages of various approaches to making federally funded data available to the public."

INTRODUCTION

The Task Force was formed primarily in response to the National Science Foundation's requirement that grant applications present data management plans. But it was understood from the start that the NSF requirement was only one example of a broader trend calling for scholars to "manage" their data. Implicitly, the questions raised were about digital objects and so this report explicitly only considers digital not physical objects. "Manage" in this context is understood to comprise 5 elements:

- organize data in a way that is useful for the purpose of the project
- identify data that should be preserved
- identify data that should be made accessible to other scholars
- ensure that data that are identified for preservation and sharing be suitably organized
- determine how those data will be preserved and shared.

The Task Force's charge is to draft a statement that would serve starting point for a discussion of how historians of science should address data management. Its purpose was not to develop formal guidelines for history of science scholars applying to for grants.

To begin, it is important to articulate how this new mandate differs from existing practice. While historians of science are not generally engaged in data preservation, they routinely share and expose data. They publish a great deal of data in the text and tables of their work; and the notes and bibliographies in their publications are designed to permit other scholars to understand and evaluate their sources and to review and replicate their research if necessary. In addition, historians of science routinely share unpublished data among colleagues through informal research networks.

Two traits distinguish the current subjects of discussion from existing practices. First, there is a much greater emphasis on scholars' responsibility to expose the actual data, not just citations to that data. Second, traditional forms of data-sharing among scholars are essentially private and personal transactions where the owner of the data retains some influence and control over the ways in which they are understood and used. In contrast, this report examines public and impersonal processes of making data available. The potential use of the data is wholly independent of and subsequent to the acts of preservation and exposure as conceived here.

The growing attention paid to data management by funding agencies reflects two additional trends. First, is the idea that agencies that fund research ought to have some say in how that research is disseminated. Second, is a growing trend in academia toward the increased sharing and broader dissemination of research data. In addition to these contexts, our report is also informed by the recognition that some historians of science are employing computational methods in their research for which the accumulation of and shared access to data is increasingly central.

This report therefore considers what data mean in the context of history of science, when they should and should not be shared, and what mechanisms exist or could be developed for their access and preservation. In the process, the report raises pertinent questions that should be considered and, perhaps, addressed by applicants for NSF grants but it does not provide a template for grant application plans. This report is intended to serve as a starting point for discussion, to identify issues and suggest possibilities.

WHAT ARE DATA?

The NSF defines data as:

> the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. This includes original data, but also "metadata" (e.g. experimental protocols, code written for statistical analyses, etc.
>
> It is acknowledged that there are many variables governing what constitutes "data," and the management of data, and each area of science has its own culture regarding data. *(*www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf*)*

This definition, written with an eye toward scientific research needs to be adapted to the diverse practices of history of science research. Historians use and create countless kinds of materials that fall under the definition of "material . . . necessary to validate research findings." Not only do historians work with a variety of material types, but the rights that historians have in respect to that material are also far from uniform. The NSF definition implicitly assumes that the user of this "recorded factual material" has full (or at least a standard set of) rights to preserve and disseminate that material. In contrast, the historian of science may be constrained in what he or she can do with the material beyond using it for the intended research purpose.

In order to explore the management requirements of the data used by historians of science, it is necessary first to articulate a categorization scheme that enables us to distinguish among types of data. The Task Force found that the following origin-based categories provided a useful way to think about data management issues for historians and has employed them throughout this report.

> 1. OWNED DATA: These data are material owned someone else. Usually, this material is found in libraries, archives, or other institutional repositories, but it may be in private hands. This is classic historical research material; it already exists and is generally findable by and accessible to the scholarly community (unless privately held).

2. COLLECTED DATA: These are original materials collected by the historian or associates. Examples of such materials include new oral histories or original photographs.

3. ANALYTICAL OR CREATED DATA: This category includes material, (e.g. research notes, tables, databases, statistical analyses) created by the historian through the analysis or manipulation of other, pre-existing materials (e.g. inventories, census data, maps, texts (including underlying files). Generally these data provide evidence for conclusions that could not be readily derived from the source material prior to manipulation.

WHY PRESERVE DATA?

These encompassing definitions of data might lead to the conclusion that every item, every annotation, every passing notion needs to be kept and preserved for the ages. Such an approach is neither feasible nor desirable. Instead, historians of science need to be selective in determining which of their data merit preservation.

In an academic research context, there are many reasons to preserve data, but in the end those reasons can all be encompassed in a single one. Data are preserved because of their potential future research value. That research could be further work by the scholar who originally collected or created the data, or it might be the work of scholars seeking to replicate or assess the research for which the data were originally collected; or the data might be used in projects that are entirely unrelated to the questions which they were originally collected to answer—questions that the original scholar might have never anticipated.

A scholar should therefore preserve data with a perceived or imaginable research value.

WHAT DATA NEED NOT BE PRESERVED?

As historians it is difficult for us to say about any given item that no-one will ever find it useful for research. However, in terms of the obligation to the research community, it is reasonable to say that material that does not directly and substantively contribute to the conclusions of a research project need not be preserved. Another way of thinking about it is to say that if the data are more revealing about the author of the research than the topic of the research they need not be preserved.

WHY SHARE/EXPOSE DATA?

There many reasons to share data, including but not limited to the following:

1. It is a longstanding academic practice that data which serve as evidence to support published research are made available so that the validity of that published work may be assessed. Often this is done within the publications themselves.

2. There is an emerging practice in portions of the academic community that scholars have an obligation to make data available in freely accessible locations for others to use for their own research. This practice is gaining in strength but is by no means universally accepted within the history of science at this time.

3. Funding agencies or employers may require it. The NSF, for example expects scholars to release their data in a "timely and rapid fashion."

In this context it is important to recall that sharing data does not necessarily entail unrestricted access or use by others. Exposed data are governed by the terms (license) under which they are made available. (See also the following Web site for a relevant statement on data-sharing from the NIH. http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm#898)

WHAT DATA DOES NOT NEED TO BE SHARED?

The National Science Foundation expectation is that grantees will release their data in a "timely and rapid" fashion. But it does not define those terms understanding that each academic discipline has its own norms. Here are a few reasons why historians of science might choose not to share data. This is not an exhaustive list.

1. Data do not need to be exposed before a scholar is done with them. That is, if sharing them might provide advantage to others in the field to detriment the scholar's own research program and career.

2. Data do not need to be exposed if they are more revealing about the scholar's practices (*e.g.* early drafts, annotations) than they are about the research topic.

3. Data do not need to be exposed if the historian believes that they can only be properly understood by someone who has undertaken the task of gathering/creating them in the first place. That is, if historical insight depends as much or more on process as it does on results.

4. Data may not be exposed if the historian lacks the legal right to do so.

5. Ethical considerations may also prevent the sharing of certain types of data.

Of course, the weight of these reasons changes over time, and so, the question may be asked in slightly different way. Once the determination has been made that a given set of data ought to be preserved, what are the conditions under which it is "timely" to make it widely available? This formulation emphasizes that it is not possible to create generalized schedules for the release of data. Timeliness is contingent on the specific research and career circumstances of the scholar who has amassed the data.

HOW DO THESE CONSIDERATIONS APPLY TO THE CATEGORIES OF HISTORY OF SCIENCE DATA?

OWNED DATA: Historians using this material are typically selecting and duplicating specific items of interest. Much of this material is formally published, generally available either on paper or in electronic form. Historians do not need to "manage" these data because existing practices of citation and representation are sufficient to meet the goals. But rare, unique or hard-to-access items in this category

should be considered candidates for management and there may be a role for scholars in their long-term preservation and exposure. However, these materials are owned by another entity and use of copied material is constrained by that institution's or individual's rules. Many scholars assume that once they have made a copy of archival material they can do whatever they like with that copy. But, in fact, such is rarely the case. Archives typically grant permission to copy materials for personal, scholarly use, not for widespread dissemination. Any actions to preserve and expose data (as opposed to citations of data) in this category must be done with the explicit (documented) consent of and in accordance with the rules of the owning institution or individual.

COLLECTED DATA: The salient characteristic of this category is that it consists new material owned by the historian. Because it is unique material of potentially high research value, there is a premium on long-term preservation and access. However, the historian's ability to expose this material may be constrained by others' rights associated with it. For example, unless explicitly waived, an interviewee has rights associated with the disposition of an oral history; similarly, owners of a photographed object (or, if a person, the subject of the photograph, him/herself) may have legal rights that constrain the use of the photograph—even if the historian took the picture. In addition to such legal issues, there may be ethical factors to take into consideration as well.

ANALYTICAL OR CREATED DATA: Historians produce large quantities of material during the course of a research project that potentially falls into this category. Much of it is simply work product that need not be formally managed, preserved or exposed. If data in this category are not clearly relevant to the conclusions reached in the research-- if for example they reveal more about the historian than about the research-- they do not need to be formally managed, preserved, or exposed. But if they directly enable the historian's inquiry or if they could be utilized to answer other questions, then they should be preserved and exposed.

ORGANIZATION OF DATA

Once it has been determined that data falls into a category where it needs to be preserved and ultimately shared, then they need to be handled with these goals in mind. There are two relevant components of the organization of data that need to be considered in this context. First is the file format in which the data exists. Second is the metadata scheme used to describe the data. For both components, historians should look for and adopt existing standards and best practices rather than seeking to develop their own.

There are many potential sources of such standards and best practices including:

1. The historian's institution
2. The repository where the data will reside
3. Professional societies such as the Society of American Archivists (for manuscript material, (http://www2.archivists.org/standards) or the Visual Resources Association, creator of the VRA Core for images (http://www.loc.gov/standards/vracore/).

The Task Force does not think it necessary or advisable for the Society to develop its own standards. It could, however, identify and recommend a set of existing standards that meet the needs of its members.

These standards would be accepted as the norm for history of science data in much the same way that the profession has adopted Chicago style as the norm for its bibliographic citations.

DATA PRESERVATION AND ACCESS

Long-term preservation of and access to data is neither a trivial task nor without cost. Sites need to be maintained, data need to be refreshed and compatibility assured with whatever applications are currently dominant. There should be some sort of legacy or succession plan to ensure that the data are maintained in the event that the primary hosting site were to fail. Hence, preservation and access are tasks that are beyond the ability of the individual scholar to ensure; instead historians must identify institutional repositories to house and preserve their data. Additionally, the diverse nature of historians' research data mean that historians may need to work with multiple repositories and that exposure of the data in an intellectually coherent fashion may of necessity occur independently from their preservation.

OWNED DATA: If the historian has created/or intends to create digital copies of owned materials that do not already exist in digital formats, he or she should offer to contribute those copies to the owning institution's digitization program, if such exists. In this scenario, the format and metadata would need to comply with the institution's program requirements. The historian should request permission to preserve and expose the data through a different repository if the owning institution does not have a digitizing program, declines the offer, or if the historian believes that the material should be joined with data from other sources at a single site.

COLLECTED DATA and ANALYTICAL OR CREATED DATA: Assuming the historian has obtained permission to do so from all rights holders, the historian should identify a suitable repository for the data. The first choice would be a repository that already has a programmatic emphasis into which the collected data fit. Second choice would be the repository (if such exists) of the scholar's home institution (if the historian is not an independent scholar). Again, data format and metadata would have to comply with the repositories' standards.

FUNDING

Preservation and access as described above are not free. Grants are time-limited, but the cost of preservation and access are ongoing, calling into question the sustainability of long-term preservation of and access to grant-funded research. Some institutions will absorb ongoing costs as part of their mission, but increasingly, institutional archives look to the donors of material to provide some money in support of the processing and maintenance of their donations. Such potential costs should be identified, if possible and factored into grant applications when appropriate.

TWO POTENTIAL INITIATIVES FOR THE HISTORY OF SCIENCE SOCIETY

Existing standards and institutions will likely satisfy most but not all of historians' of science requirements for the management, preservation and exposure of their data. There are two key functions that are not currently available and which provide opportunities for the History of Science Society to provide significant new services to its members. The Task Force believes that the following ideas illustrate gaps in the options available to historians of science for the management of their data. It is not

putting these specific initiatives forward as necessarily the best way to address these gaps. Instead, we believe they merit further assessment by the Society for their feasibility and utility.

1. PROJECT-BASED BIBLIOGRAPHIES: The model of distributed data preservation and exposure described here is less than fully satisfying intellectually. It means that many related resources (from the historian's perspective) would be dispersed among repositories and not linked. It would be a service to the community to make the connections among these resources visible in some way. To a large measure, historians do so already in the notes and references to their published work. But historians frequently gather far more data than they actually cite in their work so a comprehensive project-based bibliography of relevant citations with links to wherever they are located would be a new and valuable tool for scholars. The planned re-creation of the online, open-access version of the Isis CB (what bibliographer Stephen Weldon refers to as CB 2.0) could readily accommodate such project bibliographies, classify them according to existing schema and so make them available to researchers in an intellectually organized and relevant fashion.

2. AN HSS DATA REPOSITORY: The existing system of repositories has gaps in it of two types. First, many historians of science, especially, but not exclusively independent scholars, may not have access to suitable institutional repositories for their data. Second, for all historians there are unresolved questions how the cost of long-term preservation and exposure of data will be met.

   The centenary of *Isis* and the *Current Bibliography* is an appropriate occasion at which to consider whether the History of Science Society ought to add its own open-access data repository to the suite of services it offers the history of science community. Academia is increasingly embracing an ethic of providing unfettered access to data and there is growing attention among historians of science to computational research based on very large shared data sets. It may well be, therefore, that such a repository shall be an essential component of history of science research in the future. Such a repository would bring together (if not necessarily uniquely hold) data sets that could be constructed according to accepted standards but with the added features needed to make them particularly useful for historians of science. Indeed the society could establish a set of best practices; data models that could specify preferred file format and metadata schema for each class of data that it accepted. In so doing, it would become, ideally, a site where data sets created for one purpose could be merged and manipulated to address new questions. It might even be designed to incorporate private workspaces and tools for scholars who, when ready to do so, could expose the data simply by changing a privacy setting . Such a fully developed repository could streamline data management for historians of science and be a model for other learned societies.

   There are several ways in which the Society could potentially fund a repository; it could apply for grant funding directly; it could require that scholars who place materials in the repository request funds to do so as part of their own grant applications; it could offer institutional memberships (comparable to one of the ways PLoS generates funds) permitting free use by that institution's affiliates. Additionally, it could seek funding to offer grants of its own to independent scholars

and others who lack institutional or grant-based support for their research. Other funding models could be explored as well.

CONCLUSION

Trends both internal and external to the history of science profession make it timely to consider the management of data, its preservation and exposure, by the Society. In order to stimulate discussion within the History of Science community, this report has attempted to delineate the principal characteristics of history of science data and the issues associated therewith. In so doing, it offers guidance on the formation of a data management plans for historians of science, but it does not attempt to define or model such plans.